2017 Intelligent Systems Conference, IntelliSys 2017, 2018, vol.2018-January, pages 886-892

Spelling correction in english: Joint use of bi-grams and chunking

Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia

Abstract

© 2017 IEEE. The work presents the task of spelling correction realized in a batch mode with support of syntactic context. It uses the model of incomplete syntactic analysis, or chunking, described in Tesnière's dependencies. In order to improve the efficiency of chunking, the authors use a PoS-tagged dictionary of bi-grams. The program is written in Java; it uses UIMA framework and NLP@Cloud library. The paper presents the test results of the program execution on a collection of 100 clauses. It shows that the use of bi-grams can considerably increase the characteristics of the spelling corrector.

http://dx.doi.org/10.1109/IntelliSys.2017.8324234

Keywords

bi-grams, chunking, dependency model, English, Spelling correction, syntax, Tesnière

References

[1] Brill E., Moore R. C. (2000) An improved error model for noisy channel spelling correction. In ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, pp. 286-293. Association for Computational Linguistics.

[2] Chomsky N. (1956). Three Models for Description of Language. //IRE Trans. Informat. Theory, 1956, v. IT-2, p. 113-124.

[3] Chomsky N. (1957). Syntactic Structures. -The Hague: Mouton, 1957. (Reprint: Chomsky N. Syntacti Structures. -De Gruyter Mouton, 2002. -ISBN 3-11-017279-8).

[4] Damerau F. J. (1964) A technique for computer detection and correction of spelling errors. Communications of the ACM-7, pp. 171-176.

[5] Golding A. R., Roth D. (1999) A winnow-based approach to contextsensitive spelling correction //Machine learning.-Vol. 34.-1-3.-p. 107-130.

[6] Golding A. R., Schabes Y. (1996) Combining trigram-based and featurebased methods for context-sensitive spelling correction //Proceedings of the 34th annual meeting on Association for Computational Linguistics.-Association for Computational Linguistics,-pp. 71-78.

[7] Ivan Anisimov, Elena Makarova, Vladimir Polyakov. Chunking in Dependency Model and Spelling Correction in Russian and English. In Proceedings 2016 SAI Intelligent Systems Conference (IntelliSys), 21-22 September 2016, London, United Kingdom. 2016. Pp. 143-150.ISBN (IEEEXplore): 978-1-5090-1121-6. ISBN (USB) -978-1-5090-1665-5. DOI 978-1-5090-1121-6/16. (Available at: https://ru.scribd.com/document/326609639/IntelliSys-2016-Proceedings)

[8] Kernighan M. D., Church K. W., and Gale W. A. (1990) A spelling correction program based on a noisy channel model. In Proceedings of the 13th conference on Computational linguistics, pages 205-210. Association for Computational Linguistics.

[9] Kukich K. (1992) Techniques for automatically correcting words in texts. ACM Computing Surveys 24, pp. 377-439.

[10] Mays E., Damerau F. J., Mercer R. L. (1991) Context based spelling correction //Information Processing & Management. -Vol. 27.-5.-pp. 517-522.

[11] McIlroy M. D. (1982) Development of a Spelling List. AT&T Bell Laboratories

[12] Oflazer K. (1996) Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction //Computational Linguistics.-Vol. 22.- 1.-pp. 73-89.

[13] Ristad E. S., Yianilos P. N. (1998) Learning string-edit distance //Pattern Analysis and Machine Intelligence, IEEE Transactions on.-Vol. 20.-5.-p. 522-532.

[14] Tesniere, L. (1959). Elements of Structural Syntax (Éléments de syntaxe structural), Klincksieck, Paris. Préface by Jean Fourquet, professeur la Sorbonne. Second edition, reviewed and corrected. ISBN 2-252-02620-0. Re-edition of: Tesniere, L. (1959). Éléments de syntaxe structurale, Klincksieck, Paris. ISBN 2-252-01861-5

[15] Tesniere, L. (1988). Dependency Syntax : Theory and Practice, Albany, N.Y.: SUNY Press, 1988. 428 pp. ONLINE RESOURCES

[16] Apache Tika library https://tika.apache.org/(Accessed on December, 14, 2016).

[17] Corpus of Contemporary American English http://corpus.byu.edu/coca/(Accessed on December, 14, 2016).

[18] JFlex lexical analyzer generator http://jflex.de/(Accessed on December, 14, 2016).

[19] Liblevenshtein program library https://github.com/universalautomata/liblevenshtein-java (Accessed on December, 1, 2016).

[20] N-grams data. Corpus of Contemporary American English. http://www.ngrams.info/(Accessed on December, 15, 2016).

[21] Stanford Log-linear Part-Of-Speech Tagger http://nlp.stanford.edu/software/tagger.shtml (Accessed on December, 12, 2016).

[22] UIMA Homepage at the Apache Software Foundation https://uima.apache.org/(Accessed on December, 2, 2016).